CS 5525: Data Analytics I

B. Aditya Prakash

Lecture #1: Introduction
Lots of Data ....

- Customer Transactions
- E-commerce
- Finance
- Speech
- Automotive
- Internet / Web
- Healthcare
$600 to buy a disk drive that can store all of the world’s music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

$5 million vs. $400
Price of the fastest supercomputer in 1975 and an iPhone 4 with equal performance

235 terabytes data collected by the US Library of Congress by April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress
So What ?????

- Computers have become cheaper and more powerful, so storage is not an issue.
- There is often information “hidden” in the data that is not readily evident.
- Human analysts may take weeks to discover useful information.
- Much of the data is never analyzed at all.

We are drowning in data, but starving for knowledge!!!
Data contains value and knowledge

Prakash 2018
Data and Business

Recommended links
+79% clicks vs. randomly selected

Personalized News Interests
+250% clicks vs. editorial one-size-fits-all

Top Searches
+43% clicks vs. editor selected

Source: A. Machhanavajjhala
Red: official numbers from Center for Disease Control and Prevention; weekly
Black: based on Google search logs; daily (potentially instantaneously)

Detecting influenza epidemics using search engine query data
http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
Data and Government

http://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html

http://www.whitehouse.gov/blog/Democratizing-Data

Data and Culture

- Word frequencies in English-language books in Google’s database

Data and ____ your favorite subject

Sports

Journalism

Every year in America, 6,000 killers get away with murder. Why do police fail to solve so many homicides, even with more advanced investigative techniques? Can we stop these murderers from killing again and bring justice to their victims?
Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018.

Supply and demand of deep analytical talent by 2018 (thousand people):

- **2008 employment:** 150
- **Graduates with deep analytical talent:** 180
- **Others**: 30
- **2018 supply:** 300
- **Talent gap**: 140–190
- **2018 projected demand**: 440–490

50–60% gap relative to 2018 supply.

---

1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

How to extract value from data?

- Manipulate Data
  - CS, Domain expertise

- Analyze Data
  - Math, CS, Stat...

- Communicate your results
  - CS, Domain Expertise
Communication is important!

“The British government spends £13 billion a year on universities.”

– So?

– Try instead

http://wheredoesmymoneygo.org/bubblenetree-map.html#/~/total/education/university
What is Data Mining?

- Given lots of data
- Discover patterns and models that are:
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering

- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems
COURSE LOGISTICS
Course Information

- **Instructor**
  B. Aditya Prakash, Torgersen Hall 3160 F, [badityap@cs.vt.edu](mailto:badityap@cs.vt.edu)
  - Office Hours: TBD
  - Include string **CS 5525** in subject

- **Teaching Assistant**
  TBD
  - Office Hours: TBD

- **Class Meeting Time**
  Tuesdays, Thursdays, 11am-12:15pm, Goodwin Hall 125
Course Information

- Keeping in Touch
  Course web site
  
  http://www.cs.vt.edu/~badityap/classes/cs5525-Spr18/
  updated regularly through the semester
  – Piazza link on the website
Textbooks

- **Required**

Reference Books

- Charu C. Aggarwal: *Data Mining - The Textbook*. Springer 2015,
Prerequisites for the course

- Should enjoy the course 😊
- Basic knowledge in
  - linear algebra
  - probability and statistics
- Undergraduate coursework on algorithms
- Any programming course
- Please talk to the instructor about your background
Force-add

- Talk to me once after class
AND
- Fill-in these TWO surveys by 6pm EST today
  1. https://goo.gl/forms/yY2ffOQD6FJwU7GR2
  2. https://www.cs.vt.edu/graduate/forceadd

Password: given in class, during lecture
Important: both surveys need to be filled correctly to be considered.
Grading Policy

- 5 Homework Assignments (20%)
- 1 Mid-Term Exam (20%)
- 1 Final Exam (30%)
- 1 Final Course Project (30%)
There will be 5 homework Assignments.

Homeworks will be mostly conceptual

Some Programming will be required
- to run some data mining experiments
- Evaluate the results

You will not be asked to code existing algorithms
Exams

- Exams will be primarily based on homeworks
- There will be two exams
  - Mid-term in Mid-March
  - Final exam at the end of the semester
- These will be closed-book exams
- Absolutely no programming questions
- Mainly based on understanding of the material
Final Project

- Hands-on experience in practical data mining. Own topics are highly encouraged.

- Project topics:
  - Instructor will provide a list of potential projects immediately after the mid-term.

- Team of 2-3 students will be ideal for doing the project. The individual contributions within the project should be clearly specified.

- If you are concerned, please talk to the instructor and start the project early.
WARM-UP AND BASICS
Why do you need this course?

- To figure out some intelligent ways of handling the data
- To find out the behavior of an unseen/new data points.
- To identify some patterns in huge datasets
- To identify some abnormal/suspicious activities
- To provide guidelines to human experts - What to look for??

Data Analytics is a Fundamental Tool for Many Many Applications
Data Mining is ...

- “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data”

- “the science of extracting useful information from large data sets or databases”
  - Wikipedia.org

- More appropriate term will be ….

   Knowledge Discovery in Databases

   Visit Kdnuggets.com
Steps in Data Mining

1. Data Selection
2. Preprocessing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation

Result: Knowledge
Steps in the KDD Procedure

- Data Cleaning
  - (removal of noise and inconsistent records)
- Data Integration
  - (combining multiple sources)
- Data Selection
  - (only data relevant for the task are retrieved from the database)
- Data Transformation
  - (converting data into a form more appropriate for mining)
Steps in the KDD Procedure

- **Data Mining**
  - (application of intelligent methods in order to extract data patterns)

- **Model Evaluation**
  - (identification of truly interesting patterns representing knowledge)

- **Knowledge Presentation**
  - (visualization or other knowledge presentation techniques)
Related CS Topics

- Pattern Recognition
- Database Systems
- Artificial Intelligence
- Optimization
- Algorithms
- Machine Learning
- Visualization
- Statistics
Then, How is it unique???

- Differs from these topics because data mining typically deals with....
- Huge amounts of data
- Very high dimensional data
- More practical (useful) problems
Topics that will be covered

- Data Pre-processing
- Data Exploration
- Model Evaluation
- Classification
- Regression
- Clustering
- Anomaly Detection
- Association Analysis
Data Mining Tasks ...
Typical Data Mining Tasks are ...

- **Prediction Methods** *(You know what to look for)*
  - Use some variables to predict unknown or future values of other variables.

- **Description Methods** *(you don’t know what to look for)*
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996
Let us Categorize.....

- Classification [Predictive]
- Regression [Predictive]
- Association Rule Discovery [Descriptive]
- Clustering [Descriptive]
- Anomaly Detection [Predictive]
Classification

- **Given Input:** A collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.

- **Algorithm:** Find a *model* for class attribute as a function of the values of other attributes.

- **Final Goal:** previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the performance/accuracy of the model.
  - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.
An example of Dataset

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<table>
<thead>
<tr>
<th>Day</th>
<th>outlook</th>
<th>temperature</th>
<th>humidity</th>
<th>windy</th>
<th>play</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sunny</td>
<td>85</td>
<td>85</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td>80</td>
<td>90</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>overcast</td>
<td>83</td>
<td>86</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>rainy</td>
<td>70</td>
<td>96</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>rainy</td>
<td>68</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td>65</td>
<td>70</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>overcast</td>
<td>64</td>
<td>65</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>sunny</td>
<td>72</td>
<td>95</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>sunny</td>
<td>69</td>
<td>70</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td>75</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>sunny</td>
<td>75</td>
<td>70</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>overcast</td>
<td>72</td>
<td>90</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>overcast</td>
<td>81</td>
<td>75</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>rainy</td>
<td>71</td>
<td>91</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>
Classification models

<table>
<thead>
<tr>
<th>Day</th>
<th>outlook</th>
<th>temperature</th>
<th>humidity</th>
<th>windy</th>
<th>play</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sunny</td>
<td>85</td>
<td>85</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td>80</td>
<td>90</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>overcast</td>
<td>83</td>
<td>86</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>rainy</td>
<td>70</td>
<td>96</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>rainy</td>
<td>68</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td>65</td>
<td>70</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>overcast</td>
<td>64</td>
<td>65</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>sunny</td>
<td>72</td>
<td>95</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>sunny</td>
<td>69</td>
<td>70</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td>75</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>sunny</td>
<td>75</td>
<td>70</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>overcast</td>
<td>72</td>
<td>90</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>overcast</td>
<td>81</td>
<td>75</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>rainy</td>
<td>71</td>
<td>91</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>
Classification Task Examples

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
Collaborative Filtering: A Framework

**Items: I**

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>...</th>
<th>$i_j$</th>
<th>...</th>
<th>$i_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>3</td>
<td>1.5</td>
<td>....</td>
<td>...</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>$u_2$</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_i$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$r_{ij} = ?$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_m$</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Users: U**

**Unknown function**

$f: U \times I \rightarrow R$

The task:

Q1: Find Unknown ratings?
Q2: Which items should we recommend to this user?
Predictive Modeling: Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.

Examples:
- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Inter-cluster distances are maximized
Intra-cluster distances are minimized

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters.
Applications of Cluster Analysis

**Understanding**
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

**Summarization**
- Reduce the size of large data sets

Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.
Clustering: Application

- **Document Clustering:**
  - **Goal:** Find groups of documents that are similar to each other based on the *important terms* appearing in them.
    - Category: news, sports, weather, entertainment
    - Genre clustering: Similar styles: blogs, tweets, newswire
  - **Approach:** Identify frequently occurring terms in each document. Form a similarity measure based on the *frequencies* of different terms. Use it to cluster.
Association Analysis

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Rules Discovered:
- \{\text{Milk}\} \rightarrow \{\text{Coke}\}
- \{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}
Supermarket shelf management.

- **Goal**: To identify items that are bought together by sufficiently many customers.
- **Approach**: Process the data collected with barcode scanners to find dependencies among items.
- A classic rule --
  - If a customer buys diaper and milk, then he is very likely to buy beer.
  - So, don’t be surprised if you find six-packs stacked next to diapers!
Association Analysis: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf

Frequently Bought Together

- This item: Discourse as Data: A Guide for Analysis (Published in association with The Open University) by Prof Margaret Wetherell Paperback $53.42
- Discourse Theory and Practice: A Reader (Published in association with The Open University) by Prof Margaret Wetherell Paperback $56.98
- Constructing Grounded Theory: A Practical Guide through Qualitative Analysis (Introducing Qualitative Methods series) by Kathy Charmaz Paperback $40.46

Customers Who Bought This Item Also Bought

- Discourse Theory and Practice: A Reader... by Prof Margaret Wetherell $56.98
- Constructing Grounded Theory: A Practical Guide... by Kathy Charmaz $40.46
- Introducing Qualitative Research in Psychology by Caria Willig $55.35
- An Introduction to Discourse Analysis: Theory a... by James Paul Gee $31.56
Example: Web Mining

Web site

<table>
<thead>
<tr>
<th>User Id</th>
<th>Sequence of Pages Visited</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>/home → /home/A → /home/A/B → /home/C</td>
</tr>
<tr>
<td>0002</td>
<td>/home → /home/D → /home/D/E</td>
</tr>
<tr>
<td>0003</td>
<td>/home → /home/A → /home/C</td>
</tr>
</tbody>
</table>

Pattern: /home → /home/A → /home/C
Anomaly Detection

- Detect significant deviations from normal behavior
- Identify observations whose characteristics are significantly different from the rest of the data
  - Build a profile of legitimate transactions

Applications:
- Spam Filtering
- Credit Card Fraud Detection
- Network Intrusion Detection
Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
- Data from Multi-Sources
Evaluating Data Mining techniques

- **Predictive Accuracy** (ability of a model to predict future) or
- **Descriptive Quality** (ability of a model to find meaningful descriptions of the data, e.g. clusters)
- **Speed** (computation cost involved in generating and using the model)
- **Robustness** (ability of a model to work well even with noisy or missing data)
- **Scalability** (ability of a model to scale up well with large amounts of data)
- **Interpretability** (level of understanding and insight provided by the model)
Basics about Data

- Attributes and objects
- Types of Data Sets
- Data Quality issues
What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Attribute is a characteristic/feature/property.

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
Different attributes can be mapped to the same set of values

- Example: Attribute values for ID and age are integers
- But properties of attribute values can be different
  - ID has no limit but age has a maximum and minimum value
Types of Attributes

- **Nominal**
  - Examples: ID numbers, eye color, zip codes

- **Ordinal**
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

- **Interval**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.

- **Ratio**
  - Examples: temperature in Kelvin, length, time, counts
Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness:  \(=\neq\)
  - Order:  \(<\;>\)
  - Addition:  \(+\;-
  - Multiplication:  \(*\;/\)

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties
<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Description</th>
<th>Examples</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ((=), (\neq))</td>
<td>zip codes, employee ID numbers, eye color, sex: {male, female}</td>
<td>mode, entropy, contingency correlation, (\chi^2) test</td>
</tr>
<tr>
<td>Ordinal</td>
<td>The values of an ordinal attribute provide enough information to order objects. ((&lt;), (&gt;))</td>
<td>hardness of minerals, {good, better, best}, grades, street numbers</td>
<td>median, percentiles, rank correlation, run tests, sign tests</td>
</tr>
<tr>
<td>Interval</td>
<td>For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ((+), (-))</td>
<td>calendar dates, temperature in Celsius or Fahrenheit</td>
<td>mean, standard deviation, Pearson's correlation, (t) and (F) tests</td>
</tr>
<tr>
<td>Ratio</td>
<td>For ratio variables, both differences and ratios are meaningful. ((*), (/))</td>
<td>temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current</td>
<td>geometric mean, harmonic mean, percent variation</td>
</tr>
<tr>
<td>Attribute Level</td>
<td>Transformation</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>Any permutation of values</td>
<td>If all employee ID numbers were reassigned, would it make any difference?</td>
<td></td>
</tr>
</tbody>
</table>
| Ordinal         | An order preserving change of values, i.e., 
\[
new\_value = f(old\_value)
\]
where \( f \) is a monotonic function. | An attribute encompassing the notion of good, better best can be represented equally well by the values \{1, 2, 3\} or by \{0.5, 1, 10\}. |
| Interval        | \( new\_value = a \times old\_value + b \) 
where \( a \) and \( b \) are constants | Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree). |
| Ratio           | \( new\_value = a \times old\_value \) | Length can be measured in meters or feet. |
Discrete and Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.
Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
- Stored in sparse matrix form
- Examples:
  - Words present in documents
  - Courses taken by students
  - Items present in customer transactions
- It can be either
  - Asymmetric Binary
  - Asymmetric Discrete
  - Asymmetric Continuous
- Most students look very similar if they are compared based on the courses that they don’t take
Student-Course Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Z</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Is X more similar to Y than Z?
  - Count the difference between the two vectors
  - Give more importance to 1 than 0
  - Is the neighborhood relationship symmetric?
# Asymmetric Attributes

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>play</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>win</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document 1</strong></td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Document 2</strong></td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Document 3</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Basics about Data

- Attributes and objects
- Types of Data Sets
- Data Quality issues
Types of Data Sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data

- **Graph**
  - World Wide Web
  - Molecular Structures

- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.

<table>
<thead>
<tr>
<th>Projection of x Load</th>
<th>Projection of y Load</th>
<th>Distance</th>
<th>Load</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.23</td>
<td>5.27</td>
<td>15.22</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>12.65</td>
<td>6.25</td>
<td>16.22</td>
<td>2.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
**Document Data**

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>play</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>win</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document 1</strong></td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Document 2</strong></td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Document 3</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
Types of Data Sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data

- **Graph**
  - World Wide Web
  - Molecular Structures

- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
Graph Data

- (1) Data with Relationships among objects
  - Examples: Generic graph and HTML Links

- [Data Mining](papers/papers.html#bbbb)
- [Graph Partitioning](papers/papers.html#aaaa)
- [Parallel Solution of Sparse Linear System of Equations](papers/papers.html#aaaa)
- [N-Body Computation and Dense Linear System Solvers](papers/papers.html#ffff)
Graph Data

- (2) Data with Objects that are Graphs
  - Substructure Mining is an important area
  - E.g. Chemical Data - Benzene Molecule: \( \text{C}_6\text{H}_6 \)
Ordered Data

(1) Sequential Data - Sequences of transactions
   – Eg. People buying DVD players, buy DVDs later.

Items/Events

( A B)  (D)  (C E)
(B D)  (C)  (E)
(C D)  (B)  (A E)

An element of the sequence
Examples of Sequence Data

- Web sequence:
  `< {Homepage} {Electronics} {Digital Cameras} {Canon Digital Camera} {Shopping Cart} {Order Confirmation} {Return to Shopping} >`

- Sequence of Courses taken by a Computer Science Major:
  `< {Algorithms and Data Structures, Introduction to Operating Systems} {Database Systems, Computer Architecture} {Computer Networks, Software Engineering} {Parallel Programming} >`

- Sequence of books checked out at a library:
  - `<{Fellowship of the Ring} {The Two Towers} {Return of the King}>`
Ordered Data

(2) Sequence data – no time stamps, but order is still important. E.g. Genome data

```
GGTTCCGCCTTCAGCCCCCGCGCC
CGCAGGGCCCCGCCCCGCCGCAGTC
GAGAAGGGCCCGCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCCGCAAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```
Ordered Data

(3) Time Series data – series of some measurements taken over time

– E.g. financial Data
Ordered Data

(4) Spatio-Temporal Data

Average Monthly Temperature of land and ocean collected for a variety of geographical locations
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data
Next Week: Data Pre-processing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation